Modelling Perceptual Effects of Phonology with ASR Systems

Bing ' Er Jiang, Ewan Dunbar, Morgan Sonderegger, Meghan Clayards, Emmanuel Dupoux

Modelling Perceptual Effects of Phonology with ASR Systems

Bing'er Jiang[1,2], **Ewan Dunbar**[2,3], **Morgan Sonderegger**[1], **Meghan Clayards**[1], **Emmanuel Dupoux**[2,3]

binger.jiang@mail.mcgill.ca, ewan.dunbar@univ-paris-diderot.fr,
{morgan.sonderegger, meghan.clayards}@mcgill.ca, emmanuel.dupoux@gmail.com
[1]Department of Linguistics, McGill University, Montreal, Canada
[2]EHESS, ENS – PSL, CNRS, INRIA, Paris, France
[3]Université de Paris, LLF, CNRS, Paris, France

Abstract

This paper explores the minimal knowledge a listener needs to compensate for phonological assimilation, one kind of phonological process responsible for variation in speech. We used standard automatic speech recognition models to represent English and French listeners. We found that, first, some types of models show language-specific assimilation patterns comparable to those shown by human listeners. Like English listeners, when trained on English, the models compensate more for place assimilation than for voicing assimilation, and like French listeners, the models show the opposite pattern when trained on French. Second, the models which best predict the human pattern use contextually-sensitive acoustic models and language models, which capture allophony and phonotactics, but do not make use of higher-level knowledge of a lexicon or word boundaries. Finally, some models *overcompensate* for assimilation, showing a (super-human) ability to recover the underlying form even in the absence of the triggering phonological context, pointing to an incomplete neutralization not exploited by human listeners.

Keywords: automatic speech recognition; computational modeling; phonological assimilation; speech perception

Introduction

This paper aims to understand phonological processes in speech perception through computational modelling. It investigates how much linguistic knowledge Automatic Speech Recognition (ASR) systems can capture, focusing on the case of language-specific phonological assimilation, one widespread type of phonological process. For example, in English *green beans*, the *n* in *green* tends to be pronounced *m*, with the place of articulation assimilated to that of the following *b* (labial). While English-speaking listeners perceptually compensate for this assimilation, perceiving the *m* as *n*, listeners whose native language is French, which does not exhibit place assimilation, do not show this behaviour. We are interested in whether ASR models compensate for phonological assimilation, and if so, what kind of knowledge they use to do so.

Phonological processes, defined as a predictable sound change when the context meets certain conditions, constitute a major source of variability in speech. Studies have found that non-canonical variants constitute 27% to 75% of instances for some sounds in conversational speech (e.g. Dilley & Pitt, 2007). Indeed, while state-of-the-art ASR systems reach near-perfect performance when given clear read speech, they have a harder time when dealing with natural conversational speech. Humans, on the other hand, have no trouble processing speech with extensive variability, suggesting that they are able to perform some kind of 'inverse phonology,' mapping the variable realizations of speech sounds to their underlying representations. This makes an interesting case for cognitive modelling to explore what knowledge or capacity makes humans good at recognizing noisy speech signals. While many behavioral studies have investigated how humans process spoken language at different levels—from specific acoustic cues to understanding entire sentences—important questions remain about how these different levels of processing are integrated and interact with each other. For example, it is hard to isolate one's phonological knowledge, as it is already acquired and can not be 'undone.' Computational models allow for full control of the system, such that one can manipulate specific components to see how each change affects the final outcome, and hence quantitatively investigate the importance of the corresponding component in human cognition. The results also inform us whether or not machine learning models constructed for very specific tasks (here, ASR) can nonetheless learn generalized knowledge to represent human perception.

Theoretical Background

Consider the following example English and French utterances:

(1.1) **Viable Change (Eng):** [...] it's my *ow[m]* plan.
(1.2) **Unviable Change (Eng):** [...] it's my *ow[m]* choice.
(1.2) **No Change (Eng):** [...] it's my *ow[n]* life.
(2.1) **Viable Change (Fr):** *ro[p]* sale
(2.2) **Unviable Change (Fr):** *ro[p]* noire
(2.3) **No Change (Fr):** *ro[b]* rouge

In running speech, the [n] in *own* in an example like (1.1) is often "assimilated" by the following [p] to [m], as [m] is the labial equivalent of [n]. English listeners perceive an [n] in (1.1) (when *plan* follows), but not in (1.2), where the assimilation is not licensed. French listeners fail to show this behavior, as French does not have this specific type of assimilation. French does, however, have voicing assimilation: the voiced [b] sound in *robe sale* is pronounced as *ro[p] sale* due to assimilation to the following [s] sound, which is voiceless. French listeners show a compensation effect in these cases, while English listeners do not (Darcy et al., 2009). Several hypotheses have been proposed to explain perceptual compensation for assimilation.

Lexical Compensation treats all variations as random noise, which can be recovered using lexical or higher-order context (e.g. Marslen-Wilson & Welsh, 1978; Samuel, 2001), so *own[m]* is treated as noise which is to be recovered because 'owm' is not a possible word. This hypothesis predicts that, in the absence of a lexicon, compensation for phonological assimilation cannot happen. In this study, we explore whether computational systems without any lexicon can exhibit compensation for assimilation.

Phonetic Compensation accounts for the compensation with a low-level phonetic mechanism. Gow (2003) and Gow Im (2004) proposed that sounds that simultaneously encode two places of articulation (like the *[n/m]* in *own plan*) are parsed onto adjacent segmental positions, when the following context explains one of the places of articulation. In this case, the recovery of /n/ from [m] can be attributed to the attraction of the labial aspects of the acoustics to the following labial segment. However, this is proposed to be a language-independent process which does not account for the language-specific compensation observed. In this study, we train different models on two languages to observe whether or not there is a language asymmetry in compensation for assimilation. Moreover, as the Phonetic Compensation theory claims that purely phonetic knowledge is sufficient for compensation, we also test whether or not the following phonological context (e.g. that /p/ follows) is needed. Lastly, we investigate whether the models show any language-universal effect, where a listener compensates for little bit of a *non*-native assimilation pattern.

Language-Specific Phonological Inference treats compensation as a language-specific mechanism that undoes the effect of assimilation rules that apply during phonological planning in production. Essentially, the listener uses knowledge of production patterns to infer the underlying phoneme that has been altered due to the assimilation context. Crucially, this account relies on language specific experience with the phonological rules or patterns affecting production and on applying this knowledge 'in reverse' to compensate for them in perception. Thereforem this theory predicts that the pattern of compensation depends on the listener's language (Gaskell et al., 1995; Gaskell, 2003; Coenen et al., 2001; Gaskell & Marslen-Wilson, 1996, 1998), which accounts for the observation in Darcy et al. (2009) that French/English listeners fail to compensate for place/voicing assimilation.

Note that, while these hypotheses offer different explanations for compensation for assimilation, we do not build models to implement exactly these hypotheses. Instead, we use the hypotheses as general guidelines for implementing models with different kinds of linguistic knowledge.

Computational models

Machine learning systems have made great progress in recent years, so much so that they can compete with humans on certain tasks. Recent research has claimed that such systems that have been constructed to optimize the performance of very specific tasks can nonetheless be used as scientific models of the brain (Jozwik et al., 2019, for vision; Linzen et al., 2016, and others for NLP). Black Box NLP is a growing research area (Linzen et al., 2016) devoted to comparing machine and human processing of words and sentences. Relatively less modeling work has been done in the area of speech processing and phonology. Moreover, while previous research has built models (TRACE, McClelland & Elman, 1986; Shortlist, Norris, 1994; Bayesian cognitive models, Norris & McQueen, 2008) to account of assimilation processes, none of them are able to take raw speech as input. These models lack a level mapping acoustics to individual sounds, and, more importantly, are not directly comparable to human responses in an experimental task. In order to fully model the process of speech perception, it is important to have a model which takes exactly the same input as human listeners.

Current study: simulation of Darcy et al. (2009)

We use on ASR models to replicate the study of Darcy et al. (2009) of English place assimilation and French voicing assimilation. In this study, listeners first heard a sentence produced by a female speaker, which could be one of the three types shown in the examples above. They then heard the target word produced in isolation by a male speaker, which was the citation form, without assimilation. After hearing both stimuli, the listener decided whether the sentence contained the word they heard later. The results reflect listeners' ability to identify the same words produced in different contexts (*No Change*), detecting assimilation (*Viable Change*) and spotting 'illegal' variants (*Unviable Change*). A control test was also done by asking a different set of participants to listen to target words extracted from the stimuli sentences (later referred to as *cut-out words*). Listeners successfully restored the original phoneme for the *viable change* cases in carrier sentences, but not in cut-out conditionswhere following context was not available.

Note that all sentence stimuli in the *unviable* condition contain non-words (e.g. *owm*), which were produced deliberately by the speaker, with the purpose of creating a minimal pair in the form of complete assimilation.

Methods

This paper uses exactly the same stimuli as in Darcy et al. (2009) to compare human and machine behavior. In order to investigate what types of information a listener needs to compensate for assimilation, we use HMM-GMM models, a traditional type of ASR model. We chose this type of model as opposed to a state-of-the-art deep learning model for two reasons: 1) unlike end-to-end deep learning ASR models, HMM-GMM models have interpretable components, each corresponding to different aspects of perception/recognition, and one can hold some component stable while changing the others, which is preferable given the purpose of the study; 2) they are easier to train with relatively good results. For the purpose of the current study, we trained a set of models

on English and French respectively to represent native listeners. The **acoustic model** (AM) represents a listener's acoustic knowledge, which maps acoustics to phones. *Triphone* acoustic models are context dependent and thus can capture allophony. The **language model** (LM) represents more general knowledge of phonotactics, capturing the statistical distribution of phone sequences.

We compare different combinations of AMs and LMs of various complexity to explore the source of compensation for assimilation, that is, whether it is mostly due to the AM, LM or depends on the combination of both. We predict that a successful ASR model is likely to be a combination of good AM and LM with contextual information. A very simple AM or a very simple LM alone should not able to tackle the effects of phonological assimilation (see below). We include simple models as control conditions, to gauge the effects of more sophisticated models.

Models as ideal listeners

We use different types of HMM-GMM models to represent listeners with different kinds of knowledge of phonetics and phonology. Specifically, a listener's task is to infer the most likely sequence of phones/words given the acoustics they hear: $P(q \mid X)$ in equation (1), where q stands for the sequence of phones and X stands for the acoustics; $\hat{q}$ is the sequence of phones whose posterior probability given the observed acoustic vectors $P(q \mid X)$ is maximal.

$$\hat{q} = \arg\max_{q} P(q|X) \tag{1}$$

The equation is further broken down according to Bayesian inference to show how the acoustic model and language model jointly determine the phone posteriors. As shown in equation (2), $P(X \mid q)$ is the likelihood of the acoustics given the phone sequences, captured by the acoustic model (AM), and $P(q)$ is the prior corresponding to the probability of the phone sequences, captured by the language model (LM).

$$\hat{q} = \arg\max_{q} P(X|q)P(q) \tag{2}$$

Note that, typically, an ASR system contains a word-level LM—a model of probabilities of sequences of words—as, in most cases, the model's task is to transcribe speech to words.We build LMs which model sequences of phones, instead of words, in order to take account of the nonwords used in the experiments: a LM over real words would assign zero (or close to zero) probability to nonwords. Using a phone-level LM avoids this issue. During training, we implemented different versions of the acoustic model and language model to represent hypothetical listeners with different knowledge.

Acoustic Model We trained three types of acoustic model for mapping acoustic information to phones: 1) a monophone AM, which categorizes phones into phonemes, irrespective of the context; 2) a triphone AM, which models the phones according to the neighbouring context: if phone m occurs in two contexts a_a and b_b, then each context can be associated with a different acoustic model;3) a triphone speaker-adapted (triphone-SA) AM, which adapts to different speakers.

Language model Our phone-level language models are based on *n-grams*, which model the distribution of *n*-phone sequences. We trained four types of LMs: 1) a null (flat) LM, where the probability of the next phone is same for all phones; 2) a unigram LM, where the probability equals the frequency of individual phones; 3) a bigram LM, where the probability of the next phone is conditioned on the previous phone; 4) a trigram LM, where the probability of the next phone is conditioned on the previous two phones.

All training was done using *Abkhazia* (Schatz et al., 2016), a Kaldi-based speech recognition package (Povey et al., 2011). Training data were 46 hours in English from Librispeech for the English models and 36 hours in French from the data used for the Zero Resource Speech Challenge 2017 (Dunbar et al., 2017). Input features were Mel Frequency Cepstral Coefficients (MFCCs) with Δ and $\Delta\Delta$ extracted from the audios, with window length of 25ms and step size of 10ms.

Procedure

After training all the models, we conducted the same experiment as Darcy et al. (2009). Illustrated in Figure 1, the task for the model is to decide whether or not the sentence presented contains the same target word as the token produced in isolation by a different speaker. In particular, the model receives the sentence stimuli and does decoding over the entire sentence. The decoding process, illustrated in equation (2), can be treated as a kind of speech perception, where the model finds the best phone sequence to explain the acoustic input. We extracted the frame-level phone posteriors, i.e., the estimated posterior probability of each phone at each frame, to represent the model's 'mental representation' of perceived sounds. The same decoding was done to the target word in isolation.

After extracting the phone posteriors, we extracted the frames corresponding to the target words in the sentences and calculated the distances between the *word in carrier sentence* and *word in isolation*. The acoustic difference between the pairs was calculated using Dynamic Time Warping (DTW), while the frame-wise distance for calculating DTW was KullbackLeibler (KL) divergence, with the isolated word as the true distribution.In particular, KL divergence is a measure of how one probability distribution is different from the other reference distribution. In our case, it measures how phones are predicted differently between the pair at each frame. DTW, an algorithm for measuring the similarity between two temporal sequences which may vary in length calculates the average distance frame-level distances along a path that optimally stretches the time axes to realign the two words. The resulting distance is the difference between the pair, which is used by the model to decide whether or not the two are the same based on whether the distance is above or below a threshold value.

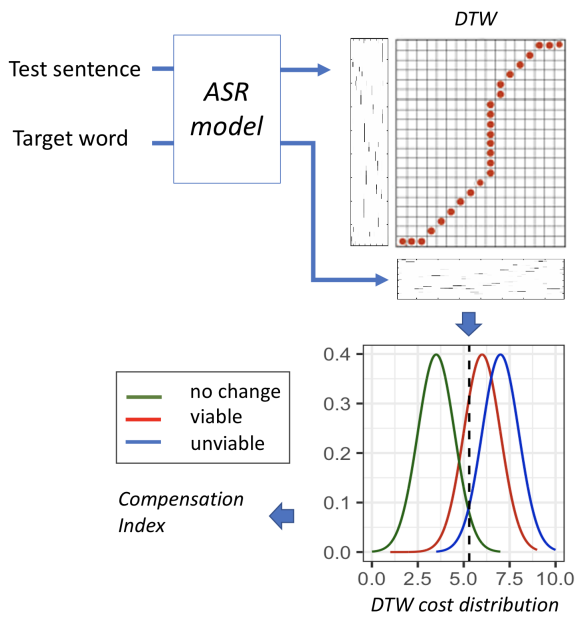

Figure 1: illustration of the pipeline

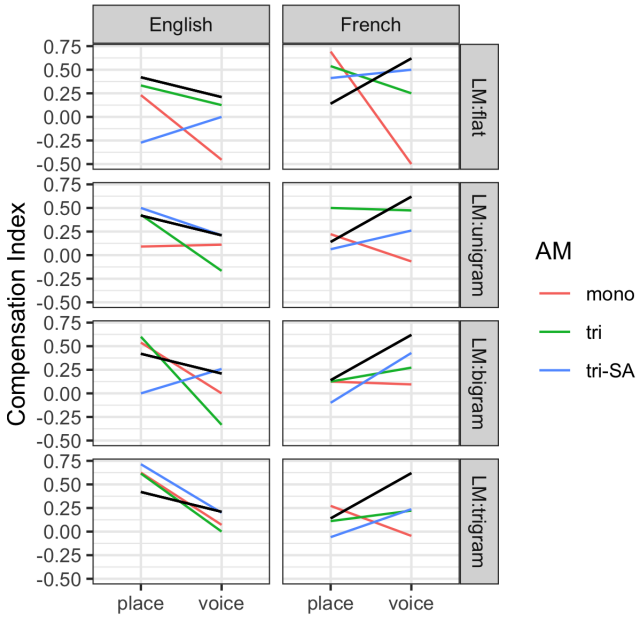

Figure 2: Compensation indices, decoded in sentence contexts across several models (colored), and humans (black).

Threshold finding and optimization We used the minimal pairs (i.e., words in *no-change* condition and *unviable* condition) to find the threshold for deciding whether or not two words are the same based on their acoustic distance (i.e. DTW cost). The threshold was then used for *viable* cases for checking whether the model is able to restore the assimilated phone.

The threshold optimization steps are illustrated in Figure 1 (bottom). The distributions represent the acoustic distances between a word *w* in carrier sentence and the same word *w'* produced in isolation by another speaker.

The algorithm searches through the range defined by the minimum distance and the maximum distance, and calculates the false negatives and false positives if it were the threshold. The threshold resulting in the smallest error is chosen. We then use the chosen threshold to determine whether a pair of phones is the same. As in the figure, distances to the left of the threshold (red line) are considered same (compensated) and distances to the right are different (uncompensated).

Evaluation: compensation index As in Darcy et al. (2009), we adopt the same compensation index for measuring the relative value of detection rate in viable condition as a function of both other conditions. As shown in the function below, the index is a ratio of "viable" to "no-change." The idea of using a ratio is to offset the perceptual biases or errors in the "unviable" condition. If all changes are detected, the index is 1; if none are detected, the index is 0.

$$Compensation\ index = \frac{(detection_{viable} - detection_{unviable})}{(detection_{no-change} - detection_{unviable})}$$

Results

Experiment 1: Sentence-level decoding

Compensation indices for models with different combinations of AM and LM are reported in Figure 2. For reference, human performance reported in Darcy et al. (2009) is shown in black. While humans display clear language-specific patterns for compensation (negative slope for English; positive slope for French), not all models are able to capture such effect. The models that best approximate human performance are triphone AM with bigram/trigram LM and triphone-SA with unigram or trigram. In particular, monophone AMs (red lines) in general fail to show the language-specific pattern, no matter which LM they pair with. They consistently compensate more for place assimilation than voice assimilation (i.e., higher Compensation Index), regardless of language. Triphone(-SA) AMs, however, can display to some extent the language-specific effect. For example, triphone AM with bigram LM shows the language-specific asymmetry: place $>$ voice / place $<$ voice for English/ French.

On the other hand, LMs (shown in rows), which represent prior knowledge of phone sequences/phonotactics, also play an important part in perceiving assimilated phones. A comparison across four types of LMs shows that one indeed needs some knowledge of phone sequences, as the flat (null) LM fails to predict the correct assimilation pattern. Other LMs all show the qualitative pattern of compensation for assimiation—different slope directions for English/French—when combined with most AMs.

Having shown that certain ASR models – those with a minimally triphone AM and a non-flat LM – do predict language-specific assimilation, we further investigate the source of compensation: how much information is in the acoustic signal? Does one need phonetic and allophonic knowledge as well as phonotactics? Can one compensate for assimilation even without the following context?

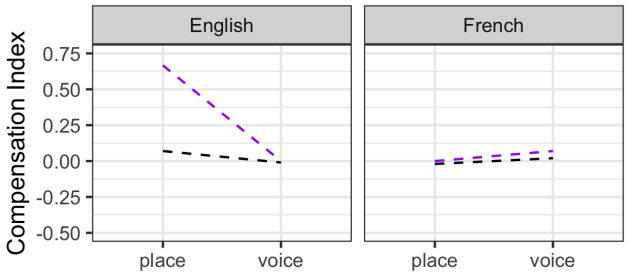

Figure 3: Compensation indices, based on raw acoustics of cut-out words (MFCCs, purple) and humans (black).

Experiment 2: Raw acoustics

In order to examine the information from raw acoustics (i.e. without using a model), we calculated the DTW distance on the MFCCs between the target words in carrier sentences and words in isolation. Note that as here the frame-level features are MFCCs, rather than phone posteriors, we use cosine distance to calculate the frame-wise distance between the pair. The rest of the procedures are the same as described above.

Figure 3 shows the compensation index for MFCCs for English and French. The purple dotted line represents the model prediction and the back dotted line is the human performance reported from the control experiment in Darcy et al. (2009).

While both model and human fail to detect assimilation from raw acoustics in the French case, the English model indeed predicts assimilation (0.66 for place and 0 for voice). Using the same asymmetry criteria as before, only half of the assimilation pattern can be accounted for, that is, only English results show an obvious negative slope, while the French results display a flat slope. Thus, raw acoustics are not sufficient for fully capturing the language-specific pattern.

The results further reveal that, first, for English, a listener can detect place assimilation just depending on the acoustics without any linguistic knowledge (if they are able to use all information carried in the signals), although humans do not (black dotted line); second, a comparison with Experiment 1 suggest that even for cases where raw acoustics are not distinguishable (i.e. French voice assimilation), the contextual phonetic knowledge and the phone distributions – knowledge learned by the model through training– enables a listener to compensate for assimilation native to their own language.

Experiment 3: Cut-out word decoding

We further did decoding only on the cut-out words, which replicates Darcy and colleagues' control experiment to explore whether or not a listener (i.e. model) is still able to compensate without the following context. Figure 4 shows the cut-out word decoding only on the four models that **successfully compensate for assimilation** in Experiment 1 (Figure 2, those displaying a negative slope for English and a positive slope for French at the same time). The logic of investigating only four, but not all, models is that, if a model fails to compensate for assimilation when given full information, then such mismatch with humans indicates that they

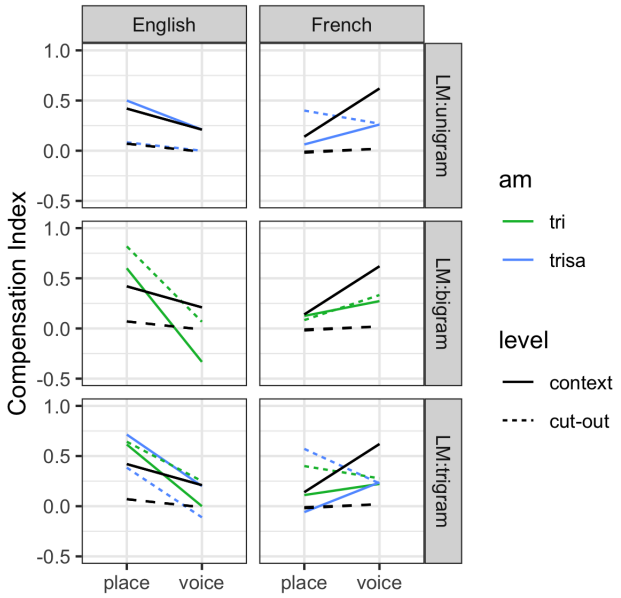

Figure 4: Compensation indices on cut-out words (dotted colored lines), compared to model performance with context (solid colored lines) and human performance with/ without context (solid/ dotted black lines). The selected models are the ones that show human-like patterns in Experiment 1.

cannot model or explain human performance, and hence are dispensed with. Dotted lines are the new results for cut-out words, while solid lines are the same results as in Figure 2, shown here for reference; black lines are human results.

Most models fail to compensate for assimilation in the absence of the following context, in that they do not show a language-specific pattern, i.e. negative slope for English and positive slope for French. The only exception is the model with triphone AM and bigram LM (middle row). It successfully show the language asymmetry of the two types of assimilation.

Discussion

This study investigates, using a computational model, whether or not a listener without the knowledge of a lexicon or word boundaries can compensate for assimilation. We replicated the psycholinguistic experiment in Darcy et al. (2009) using HMM-GMM ASR models trained on English and French corpora respectively. We found that 1) certain models do show language-specific compensation effects; 2) in most cases, though not all, following context is crucial for detecting language-specific compensation; 3) the phonetic and allophonic knowledge (captured by AM) and the distributional statistics of phones (captured by LM) both matter.

Linguistic knowledge is crucial for compensation

A major finding of this paper is that computational models trained on speech corpora indeed show language-specific compensation asymmetry like humans. The models which show this behavior share some properties: none of them use

monophone AMs and none of them use flat LMs.

While there is no successful language-specific compensation on raw acoustics, models trained on speech corpora manage to compensate for assimilation like humans (Figure 2), according to the comparison on the same experiment.

These results show the minimal knowledge a listener, whether human or machine, needs to acquire to compensate from assimilation, which cannot be done from raw acoustics alone. The successful models suggest that contextually-sensitive knowledge of both phonetics and phonology are necessary. Specifically, for phonetics, only learning phoneme category (modeled by monophone AM) is not sufficient, and one has to learn variable acoustic realizations of phones, including context information (modelled by triphone(-SA) In addition, no flat LM being successful suggests that one also needs some knowledge of phone sequences/phonotactics, but not much. The fact that even a unigram LM works (in some models) indicates that, if one is able to learn contextual phonetic information well, then it is not necessary to learn phonotactics (i.e. probabilities of phone sequences), as the relevant information is already included in the acoustic model. It also reveals that only adjacent phones matter, which is indeed the case for phonological assimilation.

The role of following context in compensation

Human listeners heavily rely on following context to detect assimilation. In order to explore whether ASR models, which already show human-like behavior, indeed show similar capacity for compensation, we test them on cut-out words where no following context is accessible (Experiment 3). The results on the four successful models show that three of them fail to compensate as humans, indicating that following context is indeed important.

On the other hand, it is interesting to note one exception, the model with triphone AM and bigram LM, which is still able to compensate without context. While this exception does not agree with Darcy et al. (2009), humans can in some cases make use of subtle acoustic cues to assimilation. In an eye-tracking study, Gow & McMurray (2007) found that for English place assimilation, listeners are able to predict the following phone (the one that triggers the assimilation) before they hear it. The over-compensating model in our study also compensates in French without following context, pointing to the availability of subtle acoustic cues not apparent in the raw MFCCs, and not exploited by humans. Nevertheless, the over-compensating model achieves greater compensation rate when taking context into consideration (difference between solid and dotted lines in Figure 4 middle row).

Acoustic signals are partially informative

The results on MFCCs show that a listener can actually detect assimilated words in English, if they are able to perceive all acoustic information. Humans, however, are not able to do this. Humans fail to perceive the full extent of acoustic detail carried in the signal, while the ASR system is optimized for solving this specific task. Nevertheless, the raw, unmodelled acoustic signals fail to identify the assimilated word in the French case.

A possible explanation for explaining the compensation effects in the cut-out stimuli could be the way we derived MFCCs: we calculated MFCCs at the sentence level and extracted frames corresponding to the word. When calculating MFCCs at each frame, a small window of signal is used for calculation. Thus, although we only extracted MFCCs up to the word boundaries, the MFCCs nevertheless contain some information about the following signal,and hence may capture some of the acoustics of the following consonant.

Conclusion

In this paper, we used ASR systems to represent listeners for modelling language-specific phonological assimilation. We found that certain models indeed reproduced human behavior, not only in what they *can* do—compensation for assimilation, but also in what they *cannot* do—no compensation for assimilation without following context. Moreover, these computational 'listeners' do not employ any of the higher-level knowledge sometimes used to explain perception of cross-word assimilation by human listeners: a lexicon, explicit phonological rules, or word boundaries. The patterns are explained by a combination of contextual acoustic modelling and phonotactic patterns, but nowhere in the system is there an application of explicit (inverse) phonological rules.

In future work, we plan to better compare ASR with theoretical accounts based on the lexical level. This can be done by using a word-level LM with a fallback on a phone-level LM for unseen words. We also plan to test more modern hybrid ASR systems based on deep neural networks for the AM (Mohamed et al., 2012). Such models would presumably have better performances than the AM used in this paper, but because neural networks incorporate more context than GMMs, they could potentially reproduce compensation for assimilation without the help of any LM.

Acknowledgments

This work was funded in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute), the CIFAR LMB program, and a grant from Facebook AI Research (Research Grant) to the last author.

References

Coenen, E., Zwitserlood, P., & Bölte, J. (2001). Variation and assimilation in german: Consequences for lexical access and representation. *Language and Cognitive Processes*, *16*(5-6), 535–564.

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. *Variation and gradience in phonetics and phonology*, *14*, 265.

Dilley, L. C., & Pitt, M. A. (2007). A study of regressive place assimilation in spontaneous speech and its implications for

spoken word recognition. *The Journal of the Acoustical Society of America*, *122*(4), 2340–2353.

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., ... Dupoux, E. (2017). The Zero Resource Speech Challenge 2017. In *2017 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 323–330).

Gaskell, M. G. (2003). Modelling regressive and progressive effects of assimilation in speech perception. *Journal of Phonetics*, *31*(3-4), 447–463.

Gaskell, M. G., Hare, M., & Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive science*, *19*(4), 407–439.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, *65*(4), 575–590.

Gow, D. W., & McMurray, B. (2007). Word recognition and phonology: The case of english coronal place assimilation. *Papers in laboratory phonology*, *9*(173-200).

Gow Jr, D. W., & Im, A. M. (2004). A cross-linguistic examination of assimilation context effects. *Journal of Memory and Language*, *51*(2), 279–296.

Jozwik, K. M., Schrimpf, M., Kanwisher, N., & DiCarlo, J. J. (2019). To find better neural network models of human vision, find better neural network models of primate vision. *BioRxiv*, 688390.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, *10*(1), 29–63.

Mohamed, A.-r., Hinton, G., & Penn, G. (2012). Understanding how deep belief networks perform acoustic modelling. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... others (2011). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding*.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–351.

Schatz, T., Bernard, M., Thiolliere, R., & Cao, X.-N. (2016). *Abkhazia* (Tech. Rep.). [Online]. Available: https://abkhazia.readthedocs.io/en/latest/index.html.